FLIP 00 PROJECT REPORT

BAOJEI ZHANG

Abstract. In this article, two things are mainly discussed. The first one: what I have learned in the recent study time and which learning tools will I use; the second one: a brief introduction to the topic

of New York City taxi fare prediction, and show the data processing and prediction process.

Contents

Date: 2020-10-10.
Key words and phrases. Application software learning, Kaggle topic selection.

1. Application

LaTex: Learn the basic syntax of latex and be able to use templates for document editing.

Smart Git: Learn to use smartgit to submit modified files to the Github repository, and clone the repository from Github to the local. And learn to use the command line to create a new local repository in git bash.

Python:Recently, I learned the basic grammar of Python through watching videos and practicing, and learned to use Python language to visualize numbers in the subsequent exercises, but for the visualization part, the application is still not proficient, and more practice is required.

2. kaggle project

The main requirement of New York City taxi fare prediction: use the taxi travel records in the three existing files to predict the taxi fare for subsequent trips.The New York City Taxi Fare Forecast Project provides three forecast files, test.csv, train.csv and sample_submission.csv

(1).train.csv - Input features and target fare_amount values for the training set (about 55M rows,and I will used about 1 million)

(2).test.csv - Input features for the test set (about 10K rows). Your goal is to predict fare_amount for each row.

(3).sample_submission.csv - a sample submission file in the correct format (columns key and fare_amount).

2.1. **Data analysis and processing.**

(1).Download relevant information from the kaggle official website and import it into the python environment.

(2).Find existing feature values, convert the value of object type,

- convert latitude and longitude into distance, and subdivide time.
  Import the file in pycharm and use the head() function to view the first five data:

- Then use the describe() function to roughly view the maximum and minimum values of the data, and do a basic work for the next step of data processing

- Finally view the data type.

- Then use the describe() function to roughly view the maximum and minimum values of the data, and do a basic work for the next step of data processing

TABLE 1. Data description

key	fare_amount	pickup_datetime	pickup_longitude
2009-06-15 17:26:21	4.5	2009-06-15 17:26:21 UTC	-73.844311
2010-01-05 16:52:16	16.9	2010-01-05 16:52:16 UTC	-74.016048
2011-08-18 00:35:00	5.7	2011-08-18 00:35:00 UTC	-73.982738
2012-04-21 04:30:42	7.7	2012-04-21 04:30:42 UTC	-73.987130
2010-03-09 07:51:00	5.3	2010-03-09 07:51:00 UTC	-73.968095
pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
40.721319	-73.841610	40.712278	1
40.711303	-73.979268	40.782004	1
40.761270	-73.991242	40.750562	2
40.733143	-73.991567	40.758092	1
40.768008	-73.956655	40.783762	1

TABLE 2. Data description

list	fare	Plongitude	Platitude	Dlongitude	Dlatitude	Pcount
count	100000	100000	100000	100000	100000	100000
mean	11.3546	-72.4946	39.9144	-72.4909	39.9190	1.6738
std	9.7167	10.6939	6.2256	10.4713	6.2134	1.3001
min	-44.9000	-736.5500	-74.0076	-84.6542	-74.0063	0.0000
25%	6.0000	-73.9920	40.7349	-73.9912	40.7341	1.0000
50%	8.5000	-73.9817	40.7527	-73.9800	40.7532	1.0000
75%	12.5000	-73.9669	40.7672	-73.9634	40.7681	2.0000
max	200.0000	40.7875	401.0833	40.8510	404.6166	6.0000

- Finally view the data type.

(3).Draw data visualization model diagrams, and deal with null or meaningless values in the data.

- Use the plt.plot() function to draw the price distribution histogram, the number of people distribution histogram, and the box plots corresponding to the price and the number of people.
- Split the time data into the parameters of year, month, day and week, and then calculate the travel distance based on the existing longitude and latitude of the pickup point and the longitude and latitude of the drop off point
- Now perform correlation analysis on the existing parameters

(4).See correlation value between the feature values and the fare, and the prediction model.

- Model the processed data

- Linear regression model
  - Use a linear regression model to fit the above features and build a model, so that the predicted value of the fare will be obtained.
- XGboost model
  - Use the XG boost model to model and predict, you can also get the desired fare forecast.

List of Todos

Jupyter Notebook,PyCharm,Latex,smartGit.